NHH

INSTITUTT FOR FORETAKSØKONOMI

DEPARTMENT OF BUSINESS AND MANAGEMENT SCIENCE

FOR 8 2015

ISSN: 1500-4066

February 2015

Discussion paper

Beta-creaming

BY

Jostein Lillestøl AND **Richard Sinding-Larsen**

Norges Handelshøyskole

NORWEGIAN SCHOOL OF ECONOMICS

Beta-creaming

Jostein Lillestøl[1]

Norwegian School of Economics

Richard Sinding-Larsen[2]

Norwegian University of Science and Technology

Jan. 27, 2015

Abstract

This paper considers sampling proportional to expected size from a partly unknown distribution. The applied context is the exploration for undiscovered resources, like oil accumulations in different deposits, where the most promising deposits are likely to be drilled first, based on some geologic size indicators ("creaming"). A size distribution within the Beta-class turns out to have nice analytical features in this context, and fits available data reasonably well, after rescaling[3]. The theoretical and practical consequences for the accumulation of knowledge on the underlying distribution based on this scheme, named Beta-creaming, are explored in some detail.

Keywords: Beta distribution, sampling proportional to size, resource estimation

[1] Department of Business and Management Science, Norwegian School of Economics, Helleveien 30, N-5045 Bergen, Norway; e-mail: jostein.lillestol@nhh.no

[2] Department of Geology and Mineral Resources Engineering, Norwegian University of Science and Technology, Sem Sælands veg 1, N-7491 Trondheim; e-mail: richard.sinding-larsen@ntnu.no

[3] We owe thanks to Kenneth C. Hood of ExxonMobil for providing data from the Gulf of Mexico.

Beta - Creaming

1. Introduction

Given undiscovered resources, say oil, in different deposits, and that the quantity varies according to a partly unknown distribution. Over time some of the deposits are drilled and the discovered quantity recorded, and subsequently used to confirm or establish the characteristics of this distribution. However, the sampling is scarce, and the most promising deposits are likely to be drilled first, based on some geologic size indicators ("creaming"). This will typically give a biased view of the underlying parent distribution. The parent distribution will typically be skewed with a long right tail, and the lognormal distribution is frequently used to represent the size variation. In order to analyze the consequences of creaming, it may be illuminating to consider PPS-sampling, i.e. sampling with probabilities proportional to (expected) size. Distributions with an unrestricted right tail, like the lognormal, do not allow this, and we will study the problem within the class of Beta-distribution instead. Although this is a distribution over the interval [0,1], it can be scaled to any interval [a, b}, and it accommodates skew distributions, which in small and moderate samples turn out to be indistinguishable from the log-normal.

There is an extensive literature on size distributions in relation to oil discovery, which may differ in emphasis: description, estimation or prediction. A bibliography of early references is given by Charpentier et.al. (1995). Our work mainly relates to estimation based on statistical models for non-random sampling. Seminal references in this area are the works by Kaufman and his associates, see Barouch and Kaufman (1967), Kaufman et. al. (1975), Andreatta et. al. (1986), summarized in Kaufman (1992). Other relevant references are Schuenemeyer and Drew (1983), as well as Meisner and Demirman (1981). Useful general references are the following books: Schuenemeyer and Drew (2011) on statistics in earth sciences, Kleiber and Kotz (2003) on size distributions and Gupta and Nadarajah (2004) on the Beta-distribution and its generalizations. Note that the Beta-distribution has been used previously in resource assessments, but in a different context, see Olea (2011).

2. The Beta-model and sampling proportional to size

The Beta(p,q) density is given by

$$f(x) \propto x^{p-1}(1-x)^{q-1} \quad 0 < x < 1$$

Here $\propto$ means proportional to, and the proportionality factor is the inverse of the Beta-integral $B(p,q) = \int_0^1 x^{p-1}(1-x)^{q-1}dx$.

The parameters p>0 and q>0 determines the shape of the distribution, and we have

$$E(X) = \frac{p}{p+q} \quad var(X) = \frac{pq}{(p+q)^2(p+q+1)} \quad mode(X) = \max(0, \frac{p-1}{p+q-2})$$

Suppose that an X=x is sampled with probability proportional to x, then the sampled X becomes Beta(p+1,q) instead of Beta(p,q). Below we will illustrate the consequences of this by two examples, one artificial and one with basis in a real case prospect distribution.

For the purpose of later use it may be useful to have formulas for the distribution parameters (p,q) in terms of expectation (E) and variance (V):

$$p = E(1-E)\left(\frac{E}{V} - \frac{1}{1-E}\right) \qquad q = E(1-E)\left(\frac{1-E}{V} - \frac{1}{E}\right)$$

and in terms of expectation (E) and mode (M), in the case of M>0:

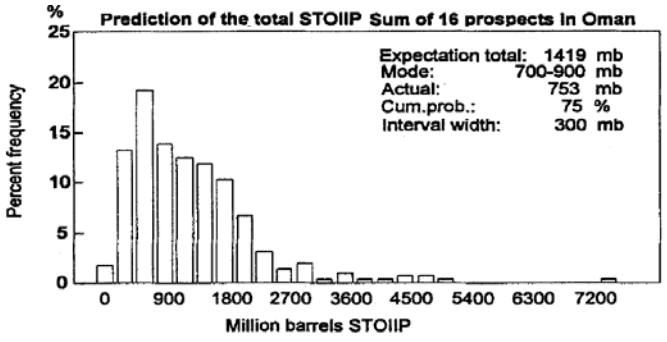
$$p = \frac{1-2M}{1-M/E} \qquad q = \frac{1-E}{E}p$$

Example 1 True Beta(2,10), sample from Beta(3,10)

Distribution	EX	Mode
True	2/12=0.167	1/10=0.100
Sampled	3/13=0.230	2/11=0.181

We see that any conclusion based on sampling, disregarding the sampling scheme, will be seriously biased.

Example 2 Prediction of total STOIIP[4]

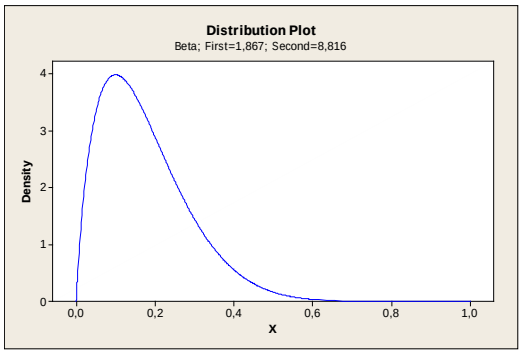


If we take the range to be [0, 8000] and scale this down to [0,1], the expectation and mode are scaled down to E=1419/8000=0.175 and M=800/8000=0.10. Using the formulas above this corresponds to the Beta-parameters p=1.867 and q=8.816, i.e. not far from those of Example 1. If this fairly represents the size distribution of the drilling opportunities, and prospects are drilled proportional to size i.e. by creaming, we therefore have learned the lesson through Example 1. [5]

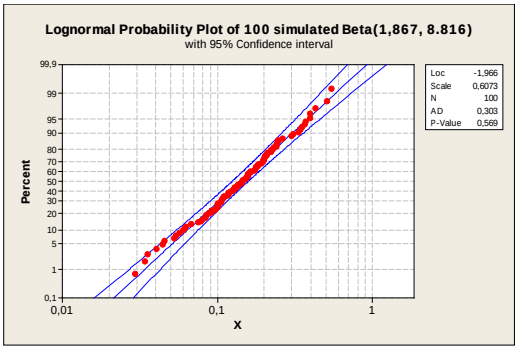
[4] STOIIP = Stock Tank Oil Initially in Place. Description of the total hydrocarbon content of a reservoir before production starts. It is distinct from 'Reserves' which can be 'recovered' or produced.

[5] This distribution is established from 16 prospects in Oman, with a log-normal size distribution for each one based on expert judgment on the mean and variance. The above distribution is then obtained by repeated simulation from each prospect and then aggregate. This is taken as the probability distribution of total STOIIP, and may itself look like a log-normal distribution.

It may be of interest to see the corresponding fitted Beta-distribution



We may also see whether it is distinguishable from the log-normal by ordinary random sampling. With a sample of n=100 we got the following lognormal probability plot, showing that log-normality cannot be rejected even with not scarce sampling.



The above results extend to sampling with probabilities proportional to any power x^d of x. Now the sampling will appear as coming from a Beta(p+d,q) distribution instead of the true B(p,q). If we sample with probability proportional to square root of size, we have to subtract ½ to get at the right distribution.

Comment. The Gamma(a,b) distribution with density

$$f(x) \propto x^{a-1}e^{-x/b} \quad x > 0$$

may seemingly share this property, with no need of downscaling to the interval [0,1]. The expectation and variance of this distribution are, respectively, EX=a·b and varX=a· b^2, which may provide some simple identification opportunities. However, this distribution has infinite support, so that any sampling proportional to power-size does not make strictly sense. Pragmatic solutions to this exist: We could truncate the distribution at some level c, or we could derive formulas as if the proportional to size distribution is proper to arrive at reduction formulas for the a-parameter exactly as above. The rationale for this is that the power expression is smeared down by the negative exponential in the product expression in the right tail.

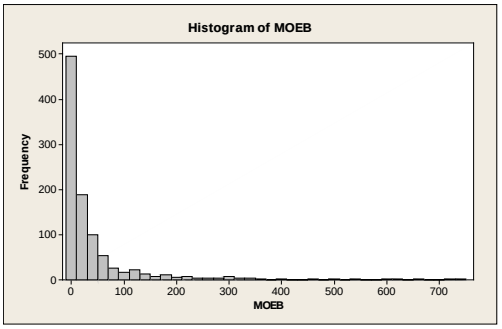
In case the Beta distribution does not fit our data, we would like to have alternatives with similar properties, i.e. power-proportional sampling keep us within the same distribution class with just a simple parameter modification. Generalizations of the Beta- family with such a feature exists, see Gupta and Nadarajah (2004). Alternatively we could go for generalizations of the Gamma distribution, based on the comment above.

The choice between the alternatives given is dependent on the balance between good fit to data, computational convenience and perhaps also to attractive and meaningful formulas. We will here explore the Beta-creaming one step further.

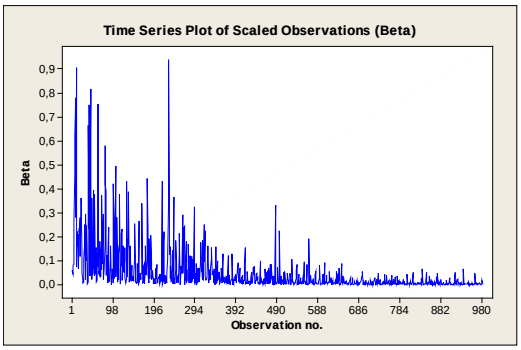
One returning issue is: How much should we degrade the computed "p"=p+d in order to get at the "true" one? This may depend on how intensively the explorers go after the large expected deposits. This may be decreasing as time goes by, so that it is natural to expect that d depends on the shape of the current expected distribution. In some contexts it may be natural to think that d must be dependent of the current p. Suppose that we observe estimated p+d, while the true one is p, and is estimated by a fraction of p+d, say by f·(p+d) which we believe to represent the true p. By setting p = f·(p+d) we get d = (1-f)/f·p. If we take f=0.75, we get d=p/3. As p decreases, so will d, so we have a device that reflects the expected exploration behavior outlined above. This device may possibly be supported by graphics from different stages of the exploration process, which also may indicate reasonable values of f.

Now to some empirics from a real exploration process, and to calculations and graphs that may reveal patterns that might turn useful for prognostic purposes in different contexts.

Example 3. The discovery sizes representing a discovery sequence from a specific geological population from the Gulf of Mexico totaling N=982 observations are shown in the histogram below. We see that the distribution of discovery sizes, with abundance of small sizes and a few outlying ones, is quite different from the one in Example2 that describes the aggregate distribution.



The data range from 0 to close to 800, and were rescaled accordingly to [0,1]. The scaled sizes in the actual discovery order revealed and recorded are given in the time series plot below. We see a pattern not consistent with i.i.d. random sampling (due to creaming), but a pattern declining over time and some other features as well.



The fit of the rescaled data to a Beta distribution is reasonably good, and we will use it as a vehicle to expose some of the opportunities and challenges we may face within a Beta-creaming framework, where the observed pattern of decreasing sizes is the result of a biased sampling scheme where the most promising prospects are likely to be drilled first.

Remarks. Before we go into the features of Beta creaming, let us briefly make some comments on the lognormal fit to this data. The time-series plot for log-sizes will show a general linear decline, possibly with some break points where the level is lifted before further decline, not so visible in the graph above. A look at the corresponding histogram for log-size will reveal a fairly good fit to the normal distribution, although it is formally rejected (P=0.006), due to too many observations in the left tail and some lack of observations in the right tail (left-skewness). It turns out that the Beta maximum likelihood fit is just slightly inferior to the corresponding lognormal. In our context it may be important to have a good fit to the extreme tails of the distribution rather than a good overall fit, and then the Beta family provides more flexibility, in particular in combination with the opportunity of estimation by quantile matching. Anyway, with as many observations as we have in this example, even small departures from any simple parametric family of distributions are likely to be picked up by formal tests and thus lead to rejection.
It is also worthwhile to note that the moment matching estimates and the (0.05, 0.95) and (0.10, 0.90) quantile matching estimates are very similar, and fit both tails heavier than maximum likelihood.

3. Some features of Beta-creaming

The Beta model may be interpreted in several contexts:

1. as an infinite population where the observations in principle may go on infinitely, i.e. biased sampling with replacement (independence),
2. as approximation to a finite population, where the sampling may go on until depletion, i.e. biased sampling without replacement (dependence),
3. as an infinite parent population where a finite random sample is taken, which is subsequently sampled without replacement (dependence).

The objective may differ: At some stage in the sampling process

a. to estimate the parameters of the distribution of remaining deposits, or
b. to estimate the parameters of the parent distribution.

The graph in Example 3 exhibit a pattern inconsistent with context 1, since we in this case necessarily will see a pattern with frequent large sizes throughout the sampling process, as if we had sampled independently from a distribution with larger expectation than the parent population, as demonstrated for the Beta-distribution above. The graph is clearly consistent with dependence, i.e. the contexts 2 and 3.

In both these contexts there are opportunities for elaborating the modeling of sequential nature. However, this may rapidly lead to complications, which make the basic issues, and thus the learning opportunities, less transparent. An obvious advantage of our Beta-scheme is that the effect of creaming is contained in the single parameter p. We will therefore explore the possibility to keep the model framework at this simple level, but for the moment leave open which is the contexts 2-3, a-b we have in mind.

To create some preliminary insight let us look at some empirics from the data of Example 3:

Having context a in mind, we consider the observation range divided in four segments Q1, Q2, Q3 and Q4 of about equal size, here with number of observations 245, 245, 245 and 247 respectively. Here are the computed means and variances and the computed values of (p+d, q) based on each segment.

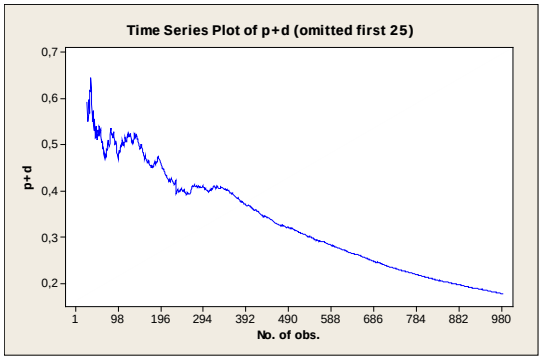
	No. of obs.	Mean	Variance	"p" = p+d	q
Q1	245	0.1319	0.0284	0.3992	2.6272
Q2	245	0.0469	0.0036	0.5381	10.9317
Q3	245	0.0161	0.0007	0.3485	21.3572
Q4	247	0.00007	0.00013	0.3870	53.1718

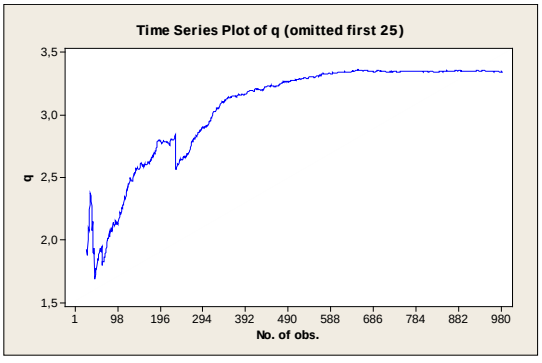
We see that p+d is about constant with no apparent monotone pattern, while q is increasing. Since p mainly affects the left tail and q the right tail, this is consistent with less frequent observations in the right tail of the distribution as we go forward. Having an estimate of (p+d,q) at a given stage in the discovery process, we need to have an idea of a representative value of d at this stage, in order to be able to estimate of the true underlying distribution of deposit sizes from which we have sampled at this stage.

Having context b in mind, we may consider computations based on enlarged segments, as follows:

	No. of obs.	Mean	Variance	"p" (p+d)	q
Q1	245	0.1319	0.0284	0.3992	2.6272
Q1+Q2	490	0.0894	0.0178	0.3198	3.3257
Q1+Q2+Q3	735	0.0649	0.0133	0.2321	3.3417
Q1+Q2+Q3+Q4	982	0.0504	0.0106	0.1775	3.3419

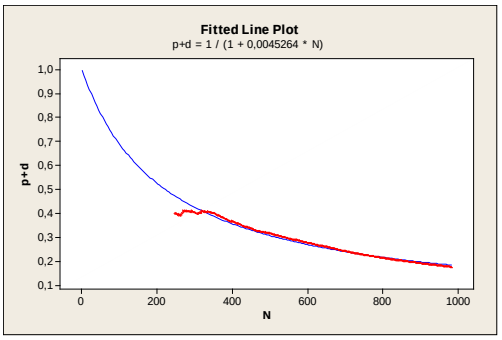
We see that the "p" decreases steadily and q increases slightly, and rapidly stabilizes after the first period, consistent with p mainly affecting the left tail and q the right tail. We do not expect much to happen in the right tail after the big fields have been sampled, but changes are still to be expected in the left tail. We could alternatively do the enlarged computations sequentially and plot the resulting p+d and q as function of the number of observations included. This gives the following (omitting the first 25 values which are very erratic due to few observations):



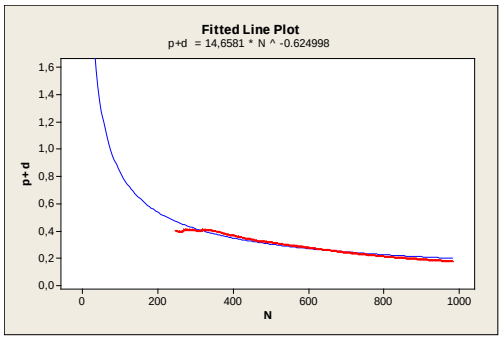


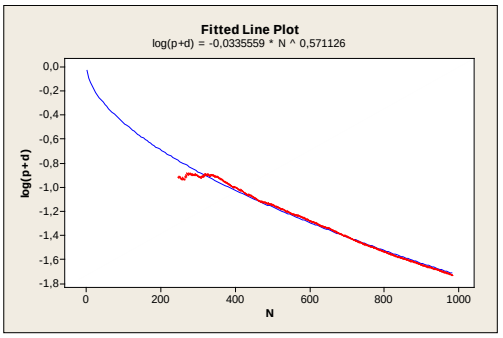
Similar plots for the first context may be based on a running window of observations.
We see a decline in p+d with n, and an apparent asymptotic behavior in q. The tail behavior of p+d appears slightly convex, but not far from linear. This tail behavior may possibly be utilized to infer the true underlying distribution in our second context.

A simple one-parameter fit to the tail of p+d, omitting the first quarter of 245 observations is given by:



Here are two two-parameter fits involving exponential dependence of with n.





We see that the log(p+d) is slightly better, and that the exponent is about 0.5, corresponding to a square-root law. A similar good fit may be obtained by log(p+d) versus log(n). Three-parameter fits including an asymptote different from zero may also be a reasonable choice, giving even better fit.

Returning to the graph of the observed sampling process, relevant questions are:

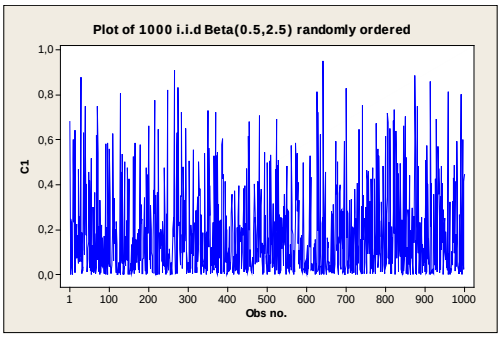
Do realistic biased sampling schemes exist that will produce this pattern?
Can we utilize the graphed behavior of (p+d, q) to say something about the parent distribution?

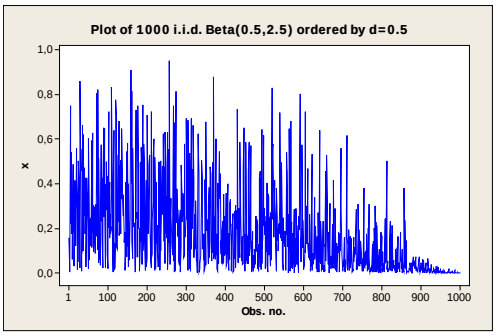
The issue is therefore: As n goes to infinity we essentially get the observations of the distribution, but we get them in an order not consistent with independent sampling (not even independent with probabilities proportional to size). A way of exploring this is to simulate randomly from Beta(p,q) and then order the observations according to size, or some principle related to size, and then see how successive enlarged segment calculations reflect the observed (tail) behavior.

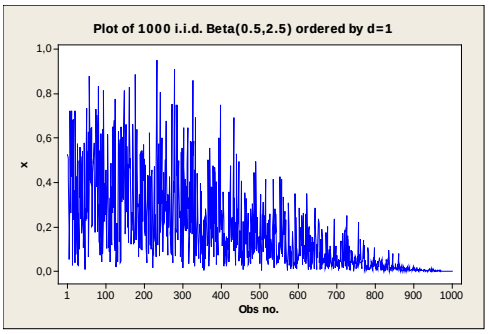
A possible context (in line with 3b above) is as follows: The geological processes which have distributed accumulation sizes in the underground above a minimum size represent a random sample of N from a continuous parent distribution, here assumed rescaled Beta. What we really observe is a sample in a discovery order where the larger sizes tend to be discovered early .

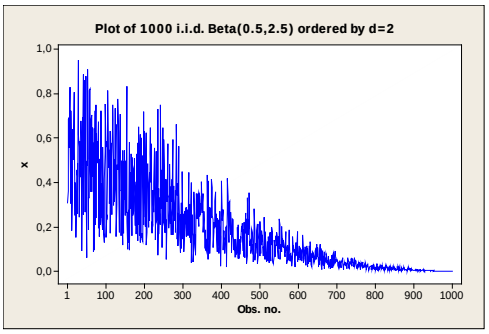
Example 4. 1000 independent Beta(0.5,2.5) are generated (top left) to constitute a population from which the 1000 items are selected without replacement, one at a time, with probabilities

proportional to x^d among the remaining items. The results are shown for d=0.5 (top right), d=1 (bottom left) and d=2 (bottom right):





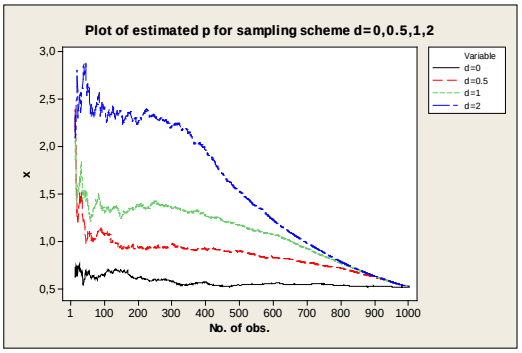


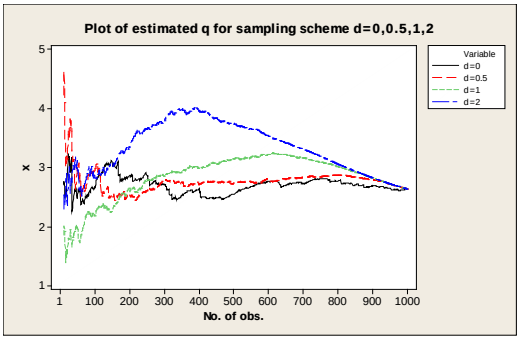


We see that the plots for d=1 and d=2 have features common with the corresponding plot for the real data example above. However, there are some differences. The existence of both high and small items for some time is more in line with d=1, but in this plot tails off too slowly. On the other hand d=2 tails off more in line with the real case, but lacks the small items in the beginning. For the real data we have both high and low ones for some time. This may reflect special features of the exploration process, among them that seismic provide better control of the large fields, in the sense that it is not that easy to avoid the small ones as it is to find the large ones. Generation of new information and practical exploration considerations may also affect the process.

If we go for a model of this kind, the choice of d is not obvious. It may seem that a mixture between d=0.5 and d=2 would give a pattern closer to the one observed, while d=1 is a compromise.

It is of interest to see how successive computations of the Beta-parameters differ in the four cases d=0, 0.5, 1, 2. In the following graph the first 10 calculated values are omitted, in order to get higher resolution in the graph, otherwise spoiled by high and erratic values.





First consider “the left tail parameter” p. We see from the four scenarios for d that we are likely to be close to p+d (“the true level of wrong p”) at 50 observations, despite some erratic behavior even beyond that. Moreover it seems that instead of the true p=0.5, we observe at this stage something close to p+d, for d=0.5, 1, 2, perhaps slightly below (which may be due to the sampled items). As d increases it starts sloping down earlier. This is as expected since, for larger d we get many large items early, making us to believe that there are few small ones in the population, corresponding to large value of p. However, as the large ones are more rapidly depleted by higher d, “our mind” changes earlier.

We see from the left graph that we approximately have slope –d for each curve , which suggests a crude formula for the computed p+d based on the n first observations, sampled from a pool of N observations, according to our proportional to power-size sampling scheme:

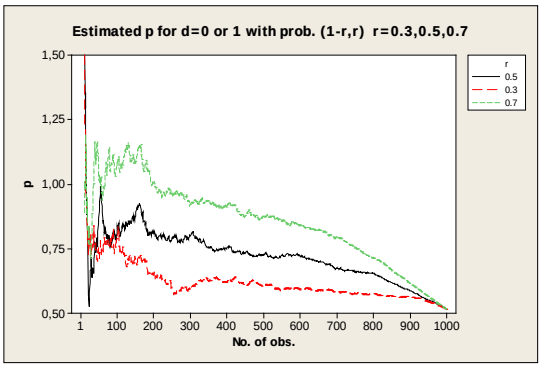
$$\text{“Computed } p+d\text{”} = f(n) = p + d(1- n/N) \text{ for } n>n(d)$$

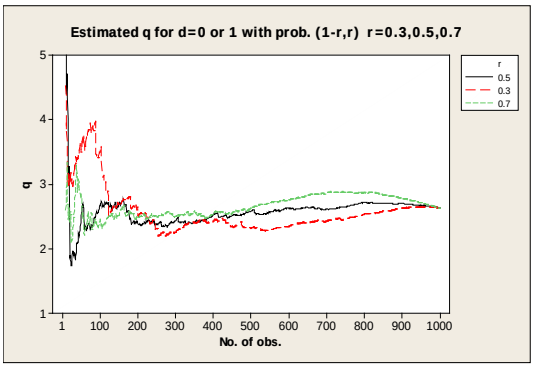
where n(d) is the threshold of necessary observations so that we are not affected by initial erratic behavior. In order to use this formula for correcting the computed p by subtracting the add on, we need an idea of d (discussed above), and how far we are in the sampling process, i.e. the fraction n/N.

Then consider the “right tail parameter q”: For q there is more uncertainty of its level in the beginning of the process. As d increases the computed q’s are eventually increasing above its true level for the entire population, reaching a peak and tailing off to a common value, when the whole population is sampled. The peak is higher and comes earlier as d increases. This is as expected, since high d leads to repeated large items in the beginning, which taken as representative of the population, lead to large computed q. In case of large d, the large items are more rapidly depleted, thus leading us to change our mind earlier. This pattern is confirmed by repeated simulations. In the plot above the effect on q is barely visible, but more visible in other simulations. If a sampling process turns out a plot with features as above, the height of the peak and point of decline is also informative of the parameter d, but this fact is more effectively observed from the plot of p+d.

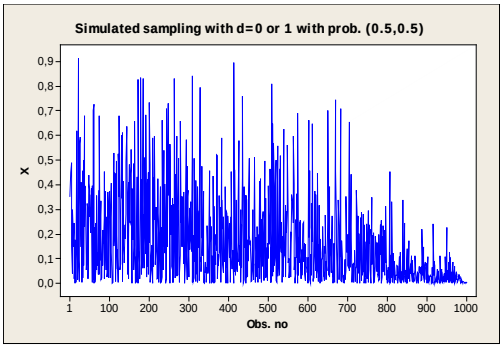
We see that the pattern for q is contrary to the asymptotic behavior in the corresponding plot for the real process in Example 3. However, we now have an explanation: We may have a mixture of two sampling processes, one with small $d=d_1$ (say zero) and one with large $d=d_2$, and then we are sure to get both small and large ones in the beginning, causing the peaks to be flattened out. We will occasionally refer to this sampling scheme [d_1: 1-r, d_2:r], so that at each instant d is randomly chosen d_1 taking value d_1 with probability 1-r and d_2 with probability r.

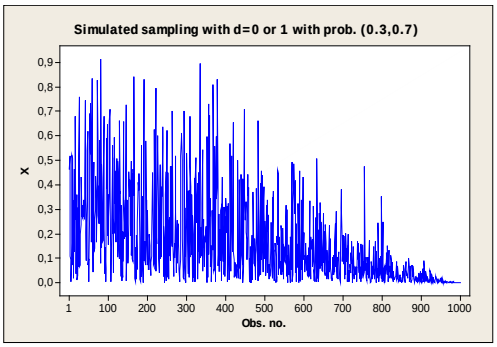
Example 4 (cont´d). We have simulated drawings from the same population of 1000 generated from Beta(0.5,2.5) as in Example 4. The drawings are according to the same scheme, but each item is sampled according to d=0 or 1 with probabilities (1-r, r) for r=0.3, 0.5, 0.7. This gives the following plots. For the p-parameter we see that we have the same downward slope as for d=1 above. For the q-parameter we have an upward slope for r=0.3, tailing off asymptotically for r=0.5, and still a peak late in the sample process for r=0.9. The pattern most consistent with the graphs for the real process of Example 3 is for r=0.5.





The graphs of the simulated sampling process is given here for r=0.5 and r=0.7. We see a slight difference in how it tapers off at the end, linearly for r=0.5 and exponential for r=0.7. This feature is supported by repeated simulations. The main difference is that the real process has both high and small sizes in the beginning, more uniformly up to a midway in the process, where we have a major drop to uniformly small sizes and not the kind of tapering off as in the graphs below. The sampling process for the real data in Example 3 seems to be more consistent with a mixture between d=0 and d>1 (say 2) until the large sizes are almost depleted, from which time the remaining occur more randomly.





From the above we have some indication that in order to explain the observed pattern we need two mixing processes with different “life lengths”, i.e. one “large size process” with short life length and a “small size process” with long life length. A possibility is then to separate the two processes. In this case this amounts to “filtering out” the contamination from the small size process in the first part of the discovery process. Then we could estimate parameters of these two processes separately. Whether this is useful may depend on the context, and there are some caveats. This approach will require that the later part of the discovery process can be used to establish the characteristics of the small size process, and so does not provide any estimates before then. For simplicity one would like to assume that the small size process is stable throughout. This may not be the case, due to depletion or learning, or a mixture of the two. If the context is that of depletion of a finite population, we have most likely found all the large accumulations and there is not much to estimate beyond

4. Opportunities for inferring population size

At this point it is worthwhile to keep in mind the context of creaming as a conscious selection of drilling opportunities according to geological indicators of best to worse sizes, which reflect the strength between the concept used for selection and the geological reality. However, this is contaminated by more or less random deviations from the correct response to the conceptual model, leading also to unexpected small discoveries early on in the discovery process.

An interesting question is whether it is possible to infer something about the size of the parent population based on the observed discoveries. We may think of two opportunities:

(i) If the recorded sizes as function of the observed number of observations start tapering off linearly we may extrapolate and take the n where we cross a low level horizontal axis as our estimate. This axis may be zero or slightly above zero, either from theoretical reasons (e.g. for an exponential decay model) or for practical reason (e.g. the small discoveries have no value)

(ii) The pattern of large discoveries may provide information, including the waiting time for maximum or an ensemble of large ones.

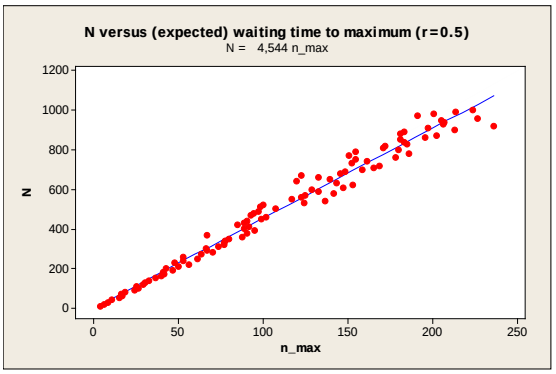
Given a biased sampling process of the kind above, the two approaches relates to the end and the beginning of the process respectively, and the most relevant (and promising) opportunity may depend on the context, among others at which stage in the sampling process we want to provide an estimate. For the first approach we may have uncovered a reasonable slope midway, which can be improved as we go along. Some of the issues related to this are explored above. For the latter approach we have to explore the consequences of the biased sampling scheme in more detail.

Comment: From other areas it is well known that precise estimation of population sizes is hard to obtain. Some areas offer specific opportunities, e.g. in wild life management by capture, tagging and recapture. Despite this transparent sampling scheme the variance of size estimates are large for reasonable efforts. In our situation the sampling scheme is less clear-cut, and consequently we may even expect less

We continue to discuss the issue within the context of a population of N accumulation sizes given as a random sample from a scaled Beta parent distribution. There are now several ways of reasoning, and one possibility is to focus on the maximum size observed in the sampling process, and utilize that the expected time will be dependent on d and N. By observing the time of the maximum it should be possible to project d and N. This may work since within any size based sampling scheme, we are fairly sure that we observe the maximum in a reasonable time compared to the size of the population. It seems hard to develop analytic formulas, and we will resort to simulations I order to establish the relationship..

Example 5. As above we take Beta(0.5,2.5) as parent population and successively simulate subsidiary populations of size N= 10, 20, 30, …, 1000. From each of these we simulate drawings according to the mixture proportional so size scheme (0,1) with probabilities (1-r, r) for r =0.0, 0.1, 0.2, …,0,9, 1.0.

For each simulation we observed the waiting time for the maximum. This was repeated 100 times for each combination of parameters, and the average $\bar{n}_{max}$ of the observed waiting times for the maximum was taken as an estimate of the expected waiting time. Here is a plot of *N* versus $\bar{n}_{max}$, for the case d=1 and r=0.5, having a clear linear structure



This simulation is then repeated for (0,d) with probabilities (1-r, r) for r =0.0, 0.1, 0.2, …,0,9, 1.0, in the case of d=1 and d=2, all showing a similar linear structure. This suggests the projection formula

$$\widehat{N} = c \cdot n_{max}$$

where, for each r, we may compute c=c(r) from regressing *N* on $\bar{n}_{max}$ based on the simulated data above. Doing this we get the following table of regression coefficients

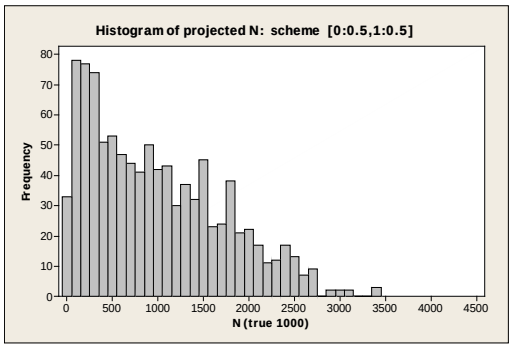
p	q	(0,d)	r=0.0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0
0.5	2.5	(0,1)	1.99	2.48	2.93	3.51	3.98	4.54	5.04	5.73	6.18	6.73	7.40
0.5	2.5	(0,2)	2.01	3.32	4.86	6.15	7.66	9.24	10.70	12.32	13.65	15.22	16.92

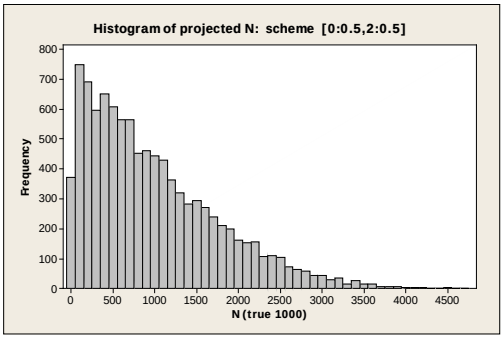
We note that the projection factor is increasing with r, and apparently also increasing with d. In the case of r=0.5, the factor is approximately double from d=1 to d=2, and slightly more than doubled beyond that. This monotonicity in the table just reflects the fact that increased probability to observe the larger fields and therefore the maximum size early will need a larger factor to project the actual population size. The table also illustrates how dependent the projections may be on the sampling scheme, and that any projection method relying on the assumption of independence is more than dubious.

We could of course have averaged more than 100 simulations to get points even closer to the regression line, but by linearity we have in effect 100·100 = 10 000 observations behind the projection formula.

In a single application the waiting time for the maximum for a given N may be far off the expectation line and so, if we after having observed an individual n_{max}, and use this for the projection of N, we may be far off as well. We illustrate this in the following:

Example 5 (cont'd): For the case of Beta(0.5,2.5) and N=1000 and sampling scheme [0: 0.5,1:0.5], for which the projection factor is 4.54, projections of N we made based on the observed n_{max} .The results of 10 000 repeats are given in the following histogram (left) and repeated for the scheme [0: 0.5,2:0.5] with projection factor 9.24 (right)





Descriptive Statistics: Projected N

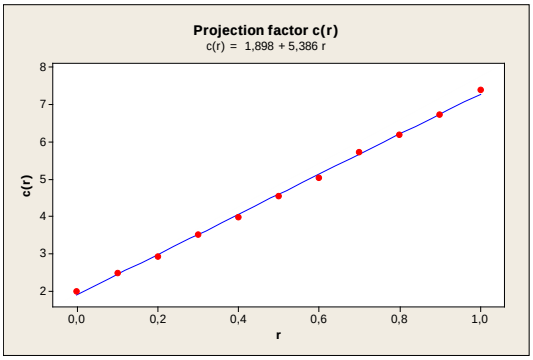
Variable	obs.	Mean	SE Mean	StDev	Minimum	Q1	Median	Q3	Max
(0,1) N	10000	978	7,18	718	4,54	378	848	1450	3538
(0,2) N	10000	977	7,75	775	9,24	360	794	1432	4749

The descriptive statistics reveal some features of interest: The mean in both cases are below 1000 by an amount (in comparison to the standard error of the mean) that indicates a systematic underestimation of N. This may be remedied by a bias correction, and our simulations so far indicate about 2% upward correction. Alternatively we may search for a more elaborate method of projection. We also see that we apparently have runs where we get the maximum at the first observation, and thus predict the population to be in the range 5-10. In practice we have to observe far beyond that in order to do any projection at all. First we may need observations to reveal the type of process, in order to choose a reasonable projection factor. Second, we got to have enough observations so that we are reasonably sure that the creamed sampling process has really revealed the maximum. This makes it obvious that some modifications are needed (if you have not realized that before!). A straightforward modification would be

$$\widehat{N} = \max(c \cdot n_{max}, n_0)$$

where n_0 is the number of observations we are having when the projection is made, or better a reasonable addition to this.

Returning to the above table for the relationship of the projection factor with r we see, by plotting the projection factor c=c(r) against r, that we have an approximate linear relationship in r as well, as shown and estimated in the following graph in the case of d=1 (note that c(0) ought to be 2, and may be forced to be so).



Discussion: We have done simulations for a specific parent population, and assumed a specific proportional to size mixture sampling scheme (0,d) for d=1,2. For this we have explored the effect of the choice of probabilities (1-r, r). We have seen that the necessary projection factor is heavily dependent on r. In the case of a comprehensive study we should of course have looked into the effect of varying p and q as well. However, we have chosen our examples to match fairly well with the real process observed in Example 3, and by this explored some of the challenges of extrapolation. For a novel process we may face a quite different sampled pattern and have to learn this as we go along, and may then be asked to make projections before we really know this. However, there is some hope that we, before being halfway through the exploration process, have obtained sufficient insight to make an educated guess. Anyway, the above may have given some insights into the limitations of projection when facing a creamed sample.

Another possibility may be to utilize the observation that the right tail of the distribution is mainly determined by the q-parameter. For independent Beta(p,q) variables we have for x close to 1 and q large enough (q>2 as it is in Example 3-4 above may be sufficient) that

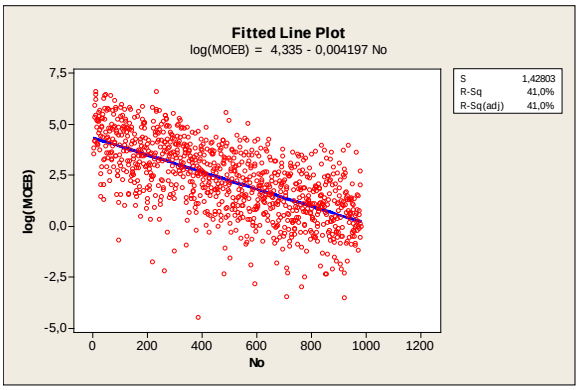
$$P(\max(x_i) \leq x) \approx 1 - n(1-x)^q$$

One may look for modifications of this that take into account our sampling scheme.

In our Gulf of Mexico data the maximum occurred at n=231. Suppose we shortly after that conclude that this is likely to be the over-all maximum. If we used the above projection method believing that we have one of the sampling schemes [0:0.5, 1:0.5] or [0:0.3, 1:0.7] the estimates of N are 1048 and 1324 respectively, i.e. not that unreasonable when we at least have come to N=982 by year 2009, and may expect discoveries beyond that. For the more realistic schemes [0:0.5, 2:0.5] and [0:0.3, 2:0.7] the estimates become a lot higher, and unrealistically so, when the following pattern is revealed. The reason for this overshoot is of course what we have noticed early on, that there are some lifts in size levels during the first half of the exploration process, maybe due to exploration focus on new play areas or new geologic of mature play areas. This means that the maximum may have a higher chance to turn up later in the process, than our simple mixture schemes will predict.

We have here investigated estimates of N by looking at maximum size, hopefully revealed early on during the exploration process, and used a specific sampling scheme as basis for revealing the underlying distribution. As stated in the beginning alternative estimates of N may be obtained by looking at the tapering off pattern at the end of the exploration process. Based on the experience gained so far this may seem more promising. However, since the tapering off pattern is revealed late in the process, the question is then how far do we have to go, in order to obtain reliable predictions.

A possibility is to plot the data on the log-scale and then do a simple linear regression fit:



We may then consider where the fitted line crosses certain low levels. This number may be taken as the "effective N" . Here we get

MOEB	0.10	0,20	0.4	0.6	0.8	1.0
LogMOEB	-2.303	-1.609	-0.916	-0.512	-0.223	0.000
Effective N	1581	1416	1251	1154	1096	1032

We see that we are likely to get about the same sloping line if we base the fit on just parts of the (abundant) data, e.g. if we do the fit early on in the exploration process. The sensitivity of the estimate of the "effective N" as to how early we do the estimate may be studied in more detail.

5. Summary and conclusions

This work has explored some of the challenges of extrapolating discovery sequences by performing simulations based on a mixed sampling proportional to size model of Beta type. The simulations match fairly well with the real exploration sequence observed from the Gulf of Mexico shelf. We may face a quite different discovery pattern in another exploration play, and will have to learn as we go along, and at the same time have to make projections before we really know how the exploration sequence behaves. However, there is some hope that before the play reaches half maturity sufficient insight is acquired to make an educated guess about the number and sizes of remaining discoveries. An advantage of using a Beta distributed parent population is that the creaming bias is represented by a single parameter that can be estimated from the discovery sequence and subsequently used for an unbiased estimation of the parent population. An important feature of the scheme is that it accounts for the abundance of small discoveries that are not well represented by common lognormal discovery models, which may significantly impact play economics. On the other hand, the choice of upper limit prior to the rescaling of observations may affect the analysis to some degree, and its sensitivity deserves a more thorough examination.

References

Andreatta, G. and Kaufman G.W. (1986) Estimation of finite population properties when sampling is without replacement and proportional to magnitude, Jour. Amer. Statist. Assoc., vol. 81, no. 395, 657-666.

Barouch, E. and Kaufman G.W. (1967) Oil and gas discovery modelled as sampling proportional to random size, *Cambridge, Mass: MIT Alfred P. Sloan School of Management.* *http://dspace.mit.edu/handle/1721.1/48701*

Charpentier, D.R., Dolton, G.L. and Ulmishek G.F. (1995) Annotated bibliography of methodology for assessment of undiscovered oil and gas resources, *Nonrenewable Resources*, vol. 4, no. 2, 154-185.

Gupta, A. K. and Nadarajah S. (2004) *Handbook of Beta Distribution and Its Applications*. New York: Marcel Dekker.

Kaufman, G.W. (1986) Finite population sampling methods for oil and gas resource estimation, in *Oil and Gas Assessment Methods and Applications (Rice D.D. ed): Amer. Assoc. Petroleum Geologists, Studies in Geology, no. 21, 43-52.*

Kaufman, G. M. (1992) Statistical Issues in the Assessment of Undiscovered Oil and Gas Resources, MIT-CEEPR-92-010WP. http://dspace.mit.edu/bitstream/handle/1721.1/50204/35719963.pdf?sequence=1

Kaufman, G. M., Balcer, Y. and Kruyt, D. (1975) A Probabilistic Model of Oil and Gas Discovery, in *Methods of Estimating the Volume of Undiscovered Oil and Gas Resources (Haun, J. D. ed), American Association of Petroleum Geologists, Studies in Geology No 1: 113-142.*

Kleiber C. and Kotz S. (2003) *Statistical Size Distributions in Economics and Actuarial Sciences*. Hoboken NJ: John Wiley and Sons Inc.

Meisner, J. and Demirmen F. (1981) The creaming method: a Bayesian procedure to forecast future oil and gas discoveries in mature exploration provinces, *Journal of the Royal Statistical Society*, v. 144, part A, p. 1-31.

Olea, R. (2011) On the use of Beta distribution in probability resource assessment, *Natural resources research,* vol. 20, no. 4, 377-388,

Schuenemeyer, J. H. and Drew L.J. (1983) A Procedure to Estimate the Parent Population of the Size Oil and Gas Fields as Revealed by a Study of Economic Truncation, *Math. Geol*. v. 15, No.1, p.145-161.

Schuenemeyer, J. H. and Drew L.J. (2011) *Statistics for Earth and Environmental Scientists.* Hoboken NJ: John Wiley and Sons Inc**.**